手术病例中结构化数据抽取研究

卢淑祺 窦志成 文继荣

(中国人民大学信息学院 北京 100872)

(中国人民大学大数据管理与分析方法研究北京市重点实验室 北京 100872)

摘 要 目前的手术病例都是以文本的方式记录的。这些文本中包含了大量对日后分析和挖掘有用的信息。通过对大量手术文本进行分析，对手术病例进行数据化和结构化，医院可以对整体病情趋势进行把握并挖掘大量对诊断有用的信息。而在针对具体病人确定手术方案时，也往往需要分析病人的历史病历，根据以前的手术情况来确定新的诊断方案。尤其对于肺部或胸腔的手术来说，确定历史手术的出血量、切除部位、切口数目以及切除范围等内容对医生制定新的手术方案具有重要意义。从历史病例中自动抽取出这些信息，将有效节省医生阅读病例的时间，进而可以让医生把更多的时间用于诊疗方案的制定上。本文重点研究胸腔手术病例中切口数量抽取问题。针对手术病例中并不直接包含切口数量、无法直接抽取的难点，本文将切口数量抽取问题转换为文本分类问题。基于文本分类的思想，首先针对病例文本中的句子着手研究，先对文本进行分句处理，选择包含切口信息的句子作为切口描述句，并基于双向 LSTM（长短期记忆神经网络）神经网络与 Attention (注意力)机制构建分句切口数目提取模型，逐个判定文本中切口描述句所记录的切口数目，最后累加切口数目。此后本文进一步构建层次化切口数目提取模型，首先针对单个句子构建双向 LSTM 网络作为句子层，并对句子层的输出再次进行过滤作为段落层的输入，构建 LSTM 神经网络作为段落层,段落层的最终输出降维得出分类结果。实验结果表明，两种切口数目判定方法准确率均可达到 98%，超出其他的多种文本分类模型如 SVM（支持向量机）以及卷积模型（TextCNN），且后者可拓展性与整体性更佳。

关键词 数目提取；文本分类；LSTM；双向 LSTM；注意力机制

中图法分类号 TP18

Research On Structural Data Extraction in Surgical Cases

LU Shu-Qi DOU Zhi-Cheng WEN Ji-Rong

(School of Information, Renmin University of China, Beijing 100872)

(Beijing Key Laboratory of Big Data Management and Analysis Methods, Renmin University of China, Beijing 100872)

Abstract Nowadays surgical cases are recorded in the form of text. These texts contain a lot of useful information for future analysis and data mining. Through the analysis of the structured information such as time, region and typical symptoms of diseases extracted from a large number of surgical texts, a hospital can grasp the overall trend of a certain kind of disease and gather much useful information for illness diagnosis. Furthermore, for a certain patient it is also necessary to analyze and understand the historical surgical records of the patient when determining the current surgical plan for him. Especial for lung and thoracic surgery, it is of great

本课题得到国家自然科学基金（No. 61872370）和国家重点基础研究发展计划/973计划(No. 2018YFC0830703)资助。卢淑祺，女，1997年生，主要研究领域为自然语言处理、数据挖掘等，E-mail:lusq@ruc.edu.cn. 窦志成（**通信作者**），男，1980年生，博士，教授，计算机学会(CCF)会员（会员号39789M），主要研究领域为信息检索、数据挖掘、大数据等，E-mail: dou@ruc.edu.cn. 文继荣，男，1972年生，博士，教授，国家千人计划专家，计算机学会(CCF)会员，主要研究领域为信息检索、数据库、大数据、数据挖掘，E-mail: jirong.wen@gmail.com.

importance and practical significance to determine the specific information such as resection site, the number of incisions and the extent of resection in the patient's historical surgical text. Automatic extraction of the structure information from historical surgical texts can save much time and effort for the doctor on reading and understanding the long surgical record, which is of great significance in the medical field. This paper focuses on automatic extraction of the number of incision in thoracic surgery cases. In order to solve the problem that the number of incisions is always not directly described in the surgical cases and cannot be directly extracted, this paper converts the problem of incision number extraction into the problem of text classification. Based on the idea of text classification, the research is first carried out on sentences in the surgical records. Firstly, the sentences describing incisions are extracted from the surgical cases (e.g. contains the word "incision"). We then judge the incision number for each sentence based on a model built on bidirectional LSTM (Long Short Term Memory) neural network and Attention mechanism. The word embedding is taken as input and the output is the prediction of incision number. Finally, the total number of incisions is counted by summing up the numbers extracted from each sentence. Considering the hierarchical relationship between words, sentences and whole text, this paper further constructs a hierarchical model suitable for incision number extraction which can be trained end-to-end. A bidirectional LSTM network along with an Attention mechanism is constructed for extracting information of each single sentence as a sentence layer, and the output of the sentence layer is filtered and then taken as the input of a paragraph layer. A LSTM neural network is constructed as a paragraph layer, which also utilizes an Attention mechanism. And the final output of the paragraph layer is then reduced in dimension to obtain the classification result. The experimental results show that the accuracy of both proposed incision number extraction model are as high as 98 %, which beyond the traditional text classification models such as SVM (Support Vector Machine) and the convolution models such as TextCNN. And the latter hierarchical incision number extraction model has better expandability and integrity, for there is need to separately predict the incision number for each sentence and sum up step by step, but is trained end to end.

Key words number extraction; text classification; LSTM; bidirectional LSTM; Attention mechanism.

1 引言

目前医院大部分是以文本报告的形式记录医疗信息，包括病人信息、医疗方案、护理措施、药品处方、手术过程等重要信息。对这些文本进行相关数据挖掘，可以增强医院的决策能力，满足医院管理与发展、临床治疗以及智能诊断的需求。例如，医院通过汇总分析手术病历报告可以得出相关规律从而优化诊治手段以及进行规划调整。李长风等人[1]通过分析处理大量基层医疗病患住院记录数据，得出基层医疗机构住院量逐年增长并存在季节性变化的结论，并由此建议合理调配区域卫生资源，提高基层住院病种诊疗防治技术。

同时，在为具体病人确定手术方案时，医院往往需要先阅读患者以往病历以及相关手术记录来确定适合的手术方案。如对于肺部的疾病，如果病人此前已有过肺部手术记录，那么很可能病人不再适合一般的手术方案。病人如果曾经在某个部位开过切口，那么再次手术时这个部位将不再适合继续作为手术对象。在这种情况下，自动挖掘手术报告中的切口数目、切口位置以及切除范围并展示给医生，将有效节省医生阅读病例的时间，而将更多的时间用于诊疗方案的研究和制定上，提升医生和医院的诊疗效率和效果。Stacey L. Slage 等人[2]曾对医师对于临床研究信息显示的感知与临床决策进行研究，调查显示结构化的临床信息相比叙述性的形式，降低了医生的认知努力程度，更能够提供临床决策支持。所以，对大量手术报告进行数据挖掘、信息抽取，对支撑医生诊断、优化手术流程、提高医院管理水平等都具有重大意义。

目前大部分医院的病例和手术报告都是以文本的形式记录和存储的。前文所述的大量对诊疗有用的信息并未直接以结构化的方式进行存储，而是蕴含在这些非结构化的文本记录中。如果依靠人工对这些病例进行关键信息的抽取，需要耗费大量人力物力。因此，利用文本挖掘技术对医疗文本进行自动数据挖掘、从无结构的病历文本中进行结构化信息抽取，是目前自然语言处理以及医疗文本分析领域的一个研究热点。近年来对于医学方面的文本数据挖掘研究也比较多样。尤其是在手术辅助、疾病预测以及智能导诊方面，对病例文本进行数据挖掘是很重要的研究手段。例如徐冉[3]针对医疗文本具有类别区分不明显、缺乏大规模训练集、某些低频词具有高判别性等特点，提出了基于 Jelinek-Merce[1]的双层 Bayes 分类模型，用于对手术文本进行病情类型分类，最后将分类器用于导诊系统中，使得病人可以远程输入症状进而自助导诊。刘利明[4]则分别基于 logistic、朴素贝叶斯以及支持向量机建立了 3 年时长的风险预警模型，能够为病患高效预测心血管疾病。

对于医疗文本信息抽取方面的研究，倪晓华[5]利用文本工程通用框架(GATE)，设计语法规则对手术病例文本进行字段抽取，一定程度上实现了病例文本的结构化，达到信息抽取、辅助治疗的效果。阮彤等人[6]也对病历文本的结构化与病历信息的标准化进行了研究，通过医疗实体识别的方式对文本中描述的症状进行抽取，达到数据结构化的目的。国外对于医疗文本中的信息抽取的相关研究也很多[7][8][9][10][11][12][13][14][15]，主要都是通过文本挖掘的手段对其中信息进行结构化。目前大多数对于医疗报告文本抽取的研究主要集中在文本中显式存在的关键信息的抽取上。比较常用的方式是通过设计基于正则表达式的规则或者基于 CRF 的实体抽取算法来实现[5]。

本文研究胸腔手术病例中的信息抽取问题。在胸腔手术病例中，描述了关于病人的大量信息，例如病人是否有吸烟史、病人吸烟史时长、病人是否曾因胸腔患病而进行过手术、病人曾经患过的胸腔疾病、病人最近一次胸腔手术日期、病人胸腔手术

表 1 示例手术病例

编号	文本	切口数量

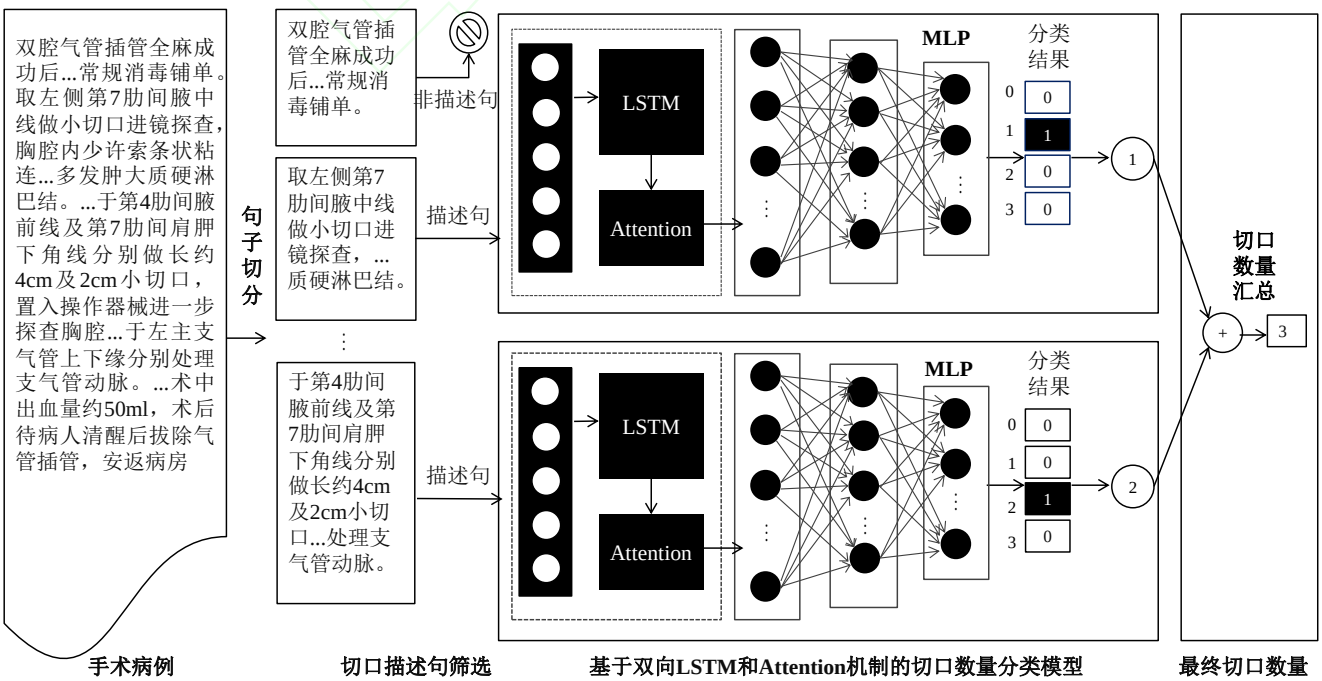

图 1：基于双向 LSTM 和 Attention 机制的手术病例分句切口数量抽取模型

中所做切口位置、切除范围、数目以及淋巴结清扫和出血量等等信息。而这些信息都将作为日后确定手术方案以及术后护理方案的依据。本文重点研究如何自动抽取出该类手术病例中的切口数量。表 1 给出了几个手术病例样本。从手术病例文本①中可以看出，对应的手术过程中一共做了三个切口（分别位于右侧第 8、5 和 7 肋间）。那么在这种情况下，确定下一次手术方案时，右侧第 8、5 和 7 肋间这些位置在术中则需要特别处理以及该范围内不再适合开新的切口。而此次手术中一共操作了三个切口，那么可以确认此次手术肋间操作密集，那么下次需要手术时手术难度将提升，需要更大的诊断治疗以及护理力度。因此切口数量识别对于确定手术方案具有重要意义。但是可以看到，在手术病例中，不存在直接的切口总数描述文字可以直接用于抽取，关于切口的信息蕴含在对于切口的描述文字中。一个病例中可以有多个句子来描述切口，并且句子可以分布在整个医疗文本的任何位置，句子的表述也不尽相同。前文所介绍的传统的基于规则的方法和基于实体抽取的方法很难适用。

针对上述问题，本文将自动抽取切口数目问题转换为文本分类问题，提出了两种基于深度学习的切口提取方法。第一种方法的结构如图 1 所示。该模型的核心思想是将文本分句，通过简单文本特征选择出描述切口的句子，采用双向 LSTM 神经网络，结合注意力机制构建句子级别的切口数目预测模型，对每个句子中描述的切口数目进行分类。最后，汇总所有句子中预测的切口数量，得出整个病例文本中的切口数目。第二种方法是直接构建整个病历文本上的端到端的切口数目分类模型，并采用两层 LSTM 分别在句子级别和病历级别学习文本向量，并最终形成分类模型，模型结构如图 3，详细将在后文第三部分进行介绍。本文的主要贡献如下：

（1）将切口数目抽取问题转换为文本分类问题。经过统计文本数据并与合作医院讨论分析得出，98%的胸腔手术病例文本中操作切口数目不超过 3 个。因此将每个候选切口描述句进行四分类，这四个类别（0、1、2、3）分别表示该句子中没有介绍新切口，一个新切口、二个新切口和三个新切口。

（2）在分类问题的基础上并不直接对病例文本进行分类，而是根据文本特点针对句子着手研究。先将整个病例按照句子进行切分，对句子进行分类后再汇总结果。

（3） 在分句研究实验的有效性上，进一步提出层次化的分类模型，一方面实现句子与文本段落分层处理，一方面实现了端到端的模型训练方法，并增强了模型的可移植性。

（4）针对分句切口抽取模型尝试多种神经网络机制，最终通过双向 LSTM 机制与注意力机制结合构建分类模型，使得分类。效果可达 98%。

（5）对于层次化分类模型，采用 LSTM、双向 LSTM 以及多层注意力机制构建分类模型，使得最终分类准确率效果达到 98.4%。

本文之后将在第二节介绍相关工作，比如医疗领域文本挖掘的相关研究、深度学习在文本挖掘中的应用；在第三节详细介绍本文提出的分句切口数目提取模型结构以及层次模型；在第四节进行实验并对实验结果进行分析，最后进行总结。

2 相关工作

2.1 文本挖掘在医疗领域的应用

随着信息技术的发展以及大数据时代的来临，文本挖掘技术被越来越多地用于各种领域，在医疗方面尤其如此。医院掌握着大量的重要数据，比如手术病例、住院记录、医生处方、用药情况等等，这些数据蕴含着医院的管理情况、资源调度、治疗手段、诊断方案等重要信息，对于优化医院管理，科学研究、辅助治疗甚至智能导诊等方面有重要的意义。但是这些数据往往不是结构化的字段，而是以文本的形式记录的，因此对于这些记录进行文本挖掘以及文本抽取具有很重要的意义。现在医疗文本挖掘、字段抽取研究以及实际应用有很多，是当下的一个研究热点。

对医疗文本进行挖掘，很重要的一个应用是训练分类模型以应用于智能诊断，通过病情描述和病例文本来自动病情分类以及病情预检。例如徐冉[1]研究了多类病例文本情况，提取文本的多级特征，基于贝叶斯分类技术构建伯努利和多项式两种病情分类模型以运用于自助导诊系统当中，实现远程自助导诊。而戴炳荣等人[16]则针对医疗卫生数据进行了分类模型构建的实验，提出了主成成分分析和支持向量机结合的分类数据挖掘方法，并在大量数据集下进行了仿真实验，有较好的分类效果。刘利明[4]也基于心血管疾病病例数据对基于支持向量机的分类模型进行了研究，并与回归模型以及贝叶斯模型进行了对比，最后将分类模型运用于检查结果

关于心血管疾病的风险预警，有较好的实验效果与实际意义。许腾[17]则针对甲状腺疾病的临床数据，提出基于随机森林的甲状腺疾病诊断结果的分类方法，从而在确定具体手术治疗方案以及用药时提供有效辅助作用。聂斌等人[18]同样通过随机森林以及人工神经网络在医疗文本分类方面进行研究，构建了糖尿病病情的初步诊断模型，在实际应用上有较为重要的意义。由于本文将切口数目抽取问题转化为分类问题，这些分类方法具有借鉴意义。

另外一个关键的应用则是通过自动化抽取电子病例文本中的字段，对其中信息进行数据化和结构化，这对于辅助治疗诊断以及医疗研究数据统计具有重要意义。倪晓华[5]利用文本工程通用框架(GATE)的应用示例组件 ANNIE 对电子病历实现首次病程记录中的自动语义标注，得到所需要的字段、句子、诊断有无危险等抽取出的信息，得出的结果符合预期的要求，有较好的实用性，可以协助医生在大段的病例文本中快速获得关键信息，从而达到辅助诊断的目的。主要方法是通过设置 GATE 规则来进行字段抽取，其中 ANNIE 提供了分词、词表查询、分句、词性标注、抽取规则定义、命名实体识别和共指消解的功能，从而实现信息抽取。阮彤等人[6]也同样在临床专病库的构建方面对病历文本的结构化与病历信息的标准化开展了研究，以文本中描述的症状为例进行了抽取实验。主要是通过抽取其中医疗实体名称及其同义词，再对识别出的症状进行语句构成成分分析，量化其中修饰词情景限定词以及程度词等，并将抽取出的症状以及检查指标与先验知识库中的症状实体进行软链接，实现症状的标准化。龚凡等人[19]则提出一种创新的基于症状构成模式的非监督学习方法来实现电子病历症状实体的抽取，主要是通过设置症状实体包含否定词修饰词等的模式组成，并基于此对病历文本进行模式识别，对识别结果进行处理得出症状列表，从而完成病历中症状信息的结构化。在生物医学方面，为了研究蛋白质关系和知识网络，林鸿飞[20]等人提出了基于支持向量机的蛋白质交互作用关系抽取方法，通过选取词项特征、关键词特征、实体距离特征等文本特征，利用 SVM 分类器判断句子中每对蛋白质是否存在相互作用关系。国外在结构化手术病例文本方面也有不少研究，例如 Joseph M Plasek 等人[7]对开发医学文本提取以及术语匹配的自然语言处理系统展开了研究，Guergana K Savova 等人[8]同样对于开发电子病历临床自由文本的信息抽取系统进行研究，主要是利用非结构化信息管理体系结构框架以及 OpenNLP 自然语言工具包对文本中的命名实体以及词性进行了识别与抽取。Maofu 等人[9]则深入研究了非结构化医学中文说明书中医学实体间语义关系的分类与提取。主要方法是根据自然语言文本的性质，从医疗说明书中提取三种文本特征，基于支持向量机构建分类模型将语义关系分类为相应的语义关系类型，最后使用提取算法可得到语义关系三元组。[10]通过启发式的文本模板生成的研究，通过在纯文本医学笔记中识别新的模板，之后在对医学文本信息提取时使用该模板进行识别，解决了包含模板的文本和文本中的模板可以被共同特征识别的问题，达到通过模板进行信息抽取的目的。[11][12]通过文本特征提取，抽取其中关于冠心病以及心力衰竭的相关判断。[13][14][15]则分别考虑从文本中抽取疾病与突变关系以及全面的药物信息，规范化这些数据以支持精确医学。

以上抽取方法主要是利用规则设计的文本匹配、模式匹配以及监督学习的办法对病例文本进行信息字段的抽取，本文在参考这些方法的基础上提出了更准确的基于深度学习的方式对术中切口字段进行信息抽取，达到量化术中切口的目的。

2.2 神经网络在文本处理中的运用

近年来，随着深度学习的兴起，关于深度学习的研究不断深入，神经网络模型越来越广泛地被应用到不同的领域，从图像处理到语音识别再到序列化语言的文本处理。神经网络发展出了很多变体，模型表现和效果也越来越好。例如 RNN 文本处理（TextRNN）以及 TextCNN[21]，都是比较基础的文本处理神经网络，但却有良好的效果，本文模型将在实验部分与这两种方法进行比较。

文献[22]中提到，微软的研究团队在成功提出深度语义结构模型之后，进一步改进模型，提出了基于单词序列的卷积深度语义结构模型，考虑到了单词的顺序信息；而为解决基于卷积神经网络无法捕捉句子长距离依存关系的问题，再提出长短时记忆的文本匹配模型 LSTM，该模型在网页搜索的在线日志数据上取得了不错的效果。

在自然语言中，一句话或者一段文本中只有部分词是关键词，包含了文本的关键信息，但在一般的循环神经网络中，一个上下文向量的产生只与其中的词向量以及顺序相关，没有考虑到词的重要性是不相同的。这个问题的解决办法最先是文献[23]

在生成式机器翻译领域提出在循环神经网络处理语句的基础上增加注意力机制，自适应地训练上下文中词的权重向量。在传统的机器翻译模型中，预测下一个词只是根据当前输入以及上一个神经元的输出计算词出现的概率。在传统的神经网络模型中，下一个词的预测值就是神经网络最后一个时刻的输出，在引入注意力机制之后，预测词的计算变为综合了此前每个时刻神经网络的输出，按照不同时刻的输出对于模型最后预测结果的不同贡献计算权重向量。最终得到词的预测值不再只是神经网络最后一个时刻的输出，而是每个时刻输出的线性加和，权重代表着模型对该部分的关注度。

文献[24]基于这个理论，提出了部分注意力模型，缩小了计算预测值参考的词的范围。这两个模型在机器翻译上面的表现都比传统的深度学习方式要好很多，使翻译的语句更加自然。而文献[25]则提出了分层注意力机制(HAN)，本文模型同样提出了分层思想，与该模型有一定的区别与联系，将在第三模型部分具体进行介绍。

3 切口数目抽取模型

给定一段手术病例文本，本文将其中的切口数量抽取问题转化为手术文本分类问题。

当需要对一段医疗文本如表 1 所示提取切口数目时，简单统计整段文本中关于切口描述的句子的数目显然不可行。首先困难在于无法判断句子描述的是一个新的切口还是前文中提到过的切口，如表 1 中的例②，其中“延长第 4 肋间切口，逐层切开胸壁组织进胸…”一句虽然是在描述切口操作，但是却是对已有切口的操作；其次无法判断一个切口描述句子中一共描述了几个切口，如表 1 中的例③，其中“于第 5 肋间腋前线及第 8 肋间肩胛下角线分别做长约 4cm 及 2cm 小切口”一句虽然只有一个“切口”，实际上描述了两个新的手术切口。在这种情况下，基于规则的抽取方法很难适用。

本文拟基于深度学习模型来解决上述问题。通过模型自动学习文本特征，构造分类模型进而提取文本中所描述的切口数目。由于文本中的切口数目一定是离散的整数，且观察文本并结合实际调研发现对于一段胸外手术描述报告，98%的手术报告切口数目不超过 3 个，所以本文基于文本分类的思想来构造分类模型。本节将分为两个部分，首先介绍分句模型，接下来基于分句模型介绍层次模型。

3.1 分句模型

本文首先将通过以下步骤处理数据，构建模型，进行实验。

（1） 数据预处理，对文本分词，进行词向量训练。

（2） 筛选切口描述句。按照训练好的词向量矩阵将切口描述句转换为词向量序列。

（3） 利用双向 LSTM 神经网络与注意力机制构建句子切口数目提取模型，对切口描述句预测其切口数目。

（4） 将切口描述句的切口数目预测结果进行累加汇总，得出整段文本的切口数目预测值。

本文接下来将在 3.1.2 节详细介绍句子的切口数目提取模型。

3.1.1 句子切分及切口描述句生成

观察文本发现如果句子要对切口进行描述，那么一定会包含“切口”一词。因此首先根据句子中有无“切口”一词筛选出切口描述句，通过训练好的词向量矩阵，将文本筛选出的描述句转换为词向量序列作为模型的输入数据。本文在此采用先分句进行切口数目预测最后再汇总句子切口数目而不是直接针对文本进行切口数目分类是为了增强模型进行切口预测的精准性。观察文本发现，一段病例文本中平均有 10-15 个完整长句，而真正描述术中切口信息的句子最多只有 4-5 个，其他的句子大多在描述有无肿液、有无粘连等术中情况。而单个句子描述相对完整，句子间相对独立，因此仅针对句子进行考虑有一定的合理性。而筛选切口描述句，并针对句子构建分类模型进行切口数目预测，相当于过滤了病例文本中的大量冗余信息，理论上能够得到更为准确的分类结果。

把含有“切口”这一关键词的句子提取出来，作为后续要处理的切口描述句。因为原始的切口数量标注是针对整个病例的，我们重新对这些切口描述句中所描述的实际切口数量进行标注。同时，对每个提取出的句子进行分词，进行词向量转换构造后面分类模型的输入。给定切口描述句 x，将其转换为：

$$\boldsymbol{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, ... \boldsymbol{x}_n\right]$$

其中 $\boldsymbol{x}_i$ 为切口描述句中第 i 个单词的词向量。

3.1.2 句子的切口数目提取模型

切口数目提取模型结构如图 2 所示。本文采用

双向 LSTM 神经网络以及注意力机制构建模型，针对句子进行切口数目预测。

如前所述，本文针对切口描述句预测其切口数目，给出的输入为词向量序列。在这种情况下，循环神经网络挖掘序列化数据中的时序信息以及语义信息的能力更为突出。但是即使是包含描述切口信息的完整句子，其中也含有很多其他的干扰信息

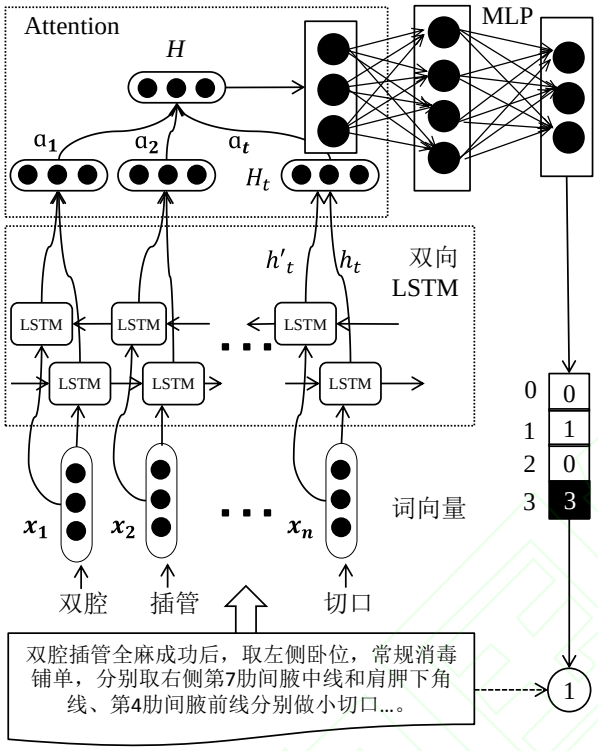

图 2 针对句子的切口提取模型

如切口处粘连表现，出血情况等等从而导致句子较长，而无关信息较多。关于切口的描述信息蕴含其中，上下文场景比较复杂，因此此时使用简单的循环神经网络容易导致模型无法保留与结尾位置相隔较远的上下文信息，性能受到限制。而 LSTM 是 RNN 的一种特殊类型，能够解决传统的 RNN 不能有效保留句子完整信息的问题。

进一步，在单个完整的句子中，语义联系紧密，并且在单个句子中判断其所描述的切口数目需要结合整个句子的信息，因为很有可能句子后半部分所描述的切口正是前半部分提到过的切口的进一步描述，为了使循环神经网络的每一个时刻可以综合上下文的信息，更好地结合上下文语境，使模型对于语义有更好的把握，本文进一步使用双向 LSTM 神经网络机制构建模型。

双向 LSTM 神经网络的基本思想就是在单个标准 LSTM 的基础上，又加了一个将信息流反向的 LSTM，从而实现提供完整的过去和未来的上下文的信息。

将句子中的词向量逐个输入双向 LSTM 层，在输出层可以得到每个时间步长神经元的正向信息流以及反向信息流的输出，将其拼接，成为最终双向 LSTM 层的输出：

$$\boldsymbol{H}_t = contact\left(\boldsymbol{h}_t , \boldsymbol{h}_t^T \right) \quad (1)$$

并将所有时间步长的输出合并可以得到矩阵：

$$\boldsymbol{K} = \left[\boldsymbol{H}_1, \boldsymbol{H}_2, \ldots \boldsymbol{H}_i, \boldsymbol{H}_{i+1}, \ldots \boldsymbol{H}_n \right]$$

在实验的过程中，如前所述观察切口描述句可以发现并不是整个句子都在描述切口，有关于切口的描述信息只是集中在一些词或短句上面。类比人工判断文本切口数目的过程中，一般会比较关注句子中对切口进行描述的部分，甚至挑选出与切口有关的句子部分单独进行分析判断，如果模型能够模拟这个过程，那么对于切口数目的判断准确度应大幅上升。如果模型能够在双向 LSTM 模型的基础上综合考虑每个时刻的输出并提高对切口描述部分的输出关注度，那么模型能够准确地找到切口描述部分并更多的参考这些部分的信息，模型的性能理应更好。基于这个考虑，参考文献[23]，在双向 LSTM 的基础上增加了注意力机制。

因此不同于一般的双向 LSTM 模型，本文的切口数目提取模型不再只处理 LSTM 层最后一个时刻的输出 $\boldsymbol{h}_n$，而是获取每个时刻的输出 $\boldsymbol{h}_i$，为每个向量训练一个权重，这个权重体现的是时刻 i 的输出向量 $\boldsymbol{h}_i$ 对结果的贡献，即模型对该时刻输出关注程度。为此，本文再构建一个神经网络层 A 来训练权重向量。

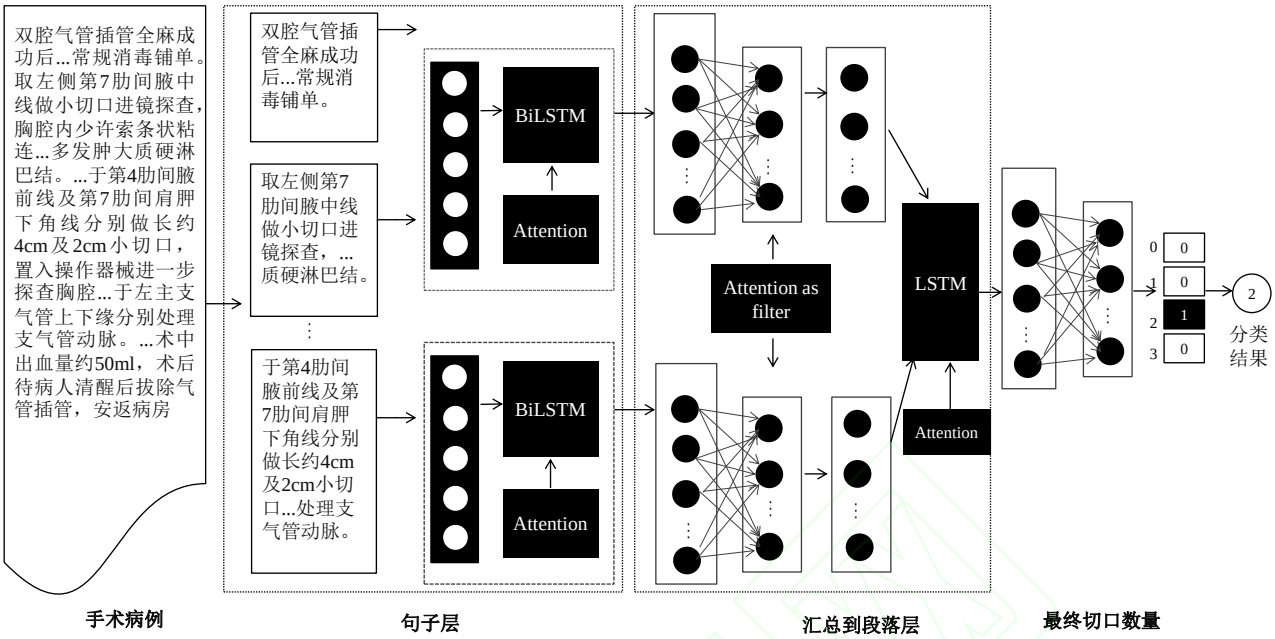

图 3：手术病例层次化切口数量抽取模型

首先，构建输入，即取之前合并各步输出的矩阵 $\boldsymbol{K}$。由于本文的目标是提取切口数目，因此在注意力层首先构建一个变量 $\boldsymbol{u}$ 作为切口信息重要性检测向量。同时为了提取各步输出中的切口信息，本文再将各步输出通过一个全连接层 A。即根据文献[23]，注意力权重值由是根据神经网络 A 中的输出与切口信息重要性检测向量 $\boldsymbol{u}$ 的相关性计算得出：

$$\boldsymbol{V} = tanh\left(\boldsymbol{K} * \boldsymbol{W}_a + \boldsymbol{b}_a^{\,T}\right) \quad (2)$$

$$\boldsymbol{t} = \boldsymbol{V} * \boldsymbol{u} \quad (3)$$

$\boldsymbol{V}$ 为 $\boldsymbol{K}$ 在神经网络 A 的输出，向量 $\boldsymbol{t}$ 则是计算 $\boldsymbol{K}$ 中每个向量中切口信息的重要程度，$\boldsymbol{t}$ 越大则重要性越强。权重向量的计算如下，即对 $\boldsymbol{t}$ 进行归一化处理：

$$e_i = exp\left(\boldsymbol{t}[i]\right) \quad (4)$$

$$\alpha_i = \frac{e_i}{\sum e_i} \quad (5)$$

最终得到向量 $\boldsymbol{\alpha}$ 就是权重向量，其中第 i 维 α_i 为 $\boldsymbol{h}_i$ 的权重。

最后求得注意力层的输出：

$$\boldsymbol{H} = \sum \boldsymbol{K}_i * \alpha_i$$

将其输入全连接层进行降维处理,为了获得分类结果再进行 softmax 处理，最终得出第 i 个切口描述句的切口数目预测值：

$$\boldsymbol{res} = \boldsymbol{W}_t * \boldsymbol{H} + \boldsymbol{b}_t \quad (6)$$

$$\boldsymbol{y}' = softmax(\boldsymbol{res}) \quad (7)$$

$$Num_i = argmax(\boldsymbol{y}') \quad (8)$$

本文基于 word2vec 训练模型得到文本中各词的词向量作为双向 LSTM 神经网络模型的输入，隐含层维度设置为 128 维，为提高速率采用批量训练，batch_size 大小设置为 32，dropout 概率设置为 0.5，通过两个全连接层降维，最后的输出维度为 4，对输出向量进行 softmax 激活函数操作，取出概率最大的序号作为每个句子分类结果。

3.1.3 切口数量计算

最后将每个切口描述句的预测值相加，得到最终对于整个文本段落的预测结果。

$$Num = \sum_{i=1}^{m} Num_i$$

3.2 层次化切口抽取模型

如前所述，针对句子级别的分类模型理论上比整段文本能够得到更为准确的结果，但是一方面增加句子级别的标注需要耗费更大的人力；另一方

面，将句子判定与最后结果加和两个步骤完全独立，难以体现模型的整体学习效果，且由于本文所用的数据中一段医疗文本中所包含的切口描述句并不多，不可避免地具有一定的数据依赖。除此以外，仅简单地筛选含“切口”的句子作为切口候选句，也具有一定的数据依赖，不具备较好的拓展性。因此本文综合考虑各方面优劣，提出层次化的切口数目提取模型，一方面结合了针对句子分类的思想，并采用了在分句实验中效果最好的双向 LSTM 与 attention 机制，另一方面构建了一个端到端的模型，保证其整体的学习能力，并且无需再次进行句子级别的标注以及候选句的筛选，保证其可拓展性。因此本文将通过如下步骤进行构建模型，进行实验。

（1） 数据预处理，进行分词以及词向量训练。

（2） 构建层次化切口数目提取模型，将医疗文本转化为词向量矩阵作为输入，输出即为医疗文本的切口数目预测结果。

接下来本文将在 3.2.1 节详细介绍层次化切口数目提取模型。

3.2.1 层次化切口数目提取模型

层次化切口数目提取模型如图 3 所示，输入为整段医疗文本，在句子层将输入的文本以句号切割，使得整段医疗文本转化为句子列表。

$$\boldsymbol{P} = \left[\boldsymbol{S}_1, \boldsymbol{S}_2, ...\boldsymbol{S}_i, \boldsymbol{S}_{i+1}, ... \boldsymbol{S}_n\right]$$

而其中，每个句子由词向量构成。

$$\boldsymbol{S}_i = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, ... \boldsymbol{x}_n\right]$$

如前所述，对于挖掘单个完整句子中的语义信息，双向 LSTM 神经网络具有优势。因此在句子层，为每个句子构建双向 LSTM 层并增加注意力机制，得到每个句子经过双向 LSTM 层处理后的输出向量。值得一提的是，不同于 HAN[25]，本文需要提取的是整段手术文本中的切口数目信息，这只是文本包含的大量信息中的一部分，因此本文关注的并不是提取能够表征完整信息的语义向量，而是需要进一步挖掘完整语义向量中有意义的部分。因此为了能够更完整地把握语义信息，这里添加注意力机制，区别于 HAN[25]，获得每个时刻输入的权重之后，并不是简单地将每个时刻地输出加权求和，而是将每个时刻的输出乘以其权重后拼接，再通过全连接层降维，进一步深度挖掘得到句子的语义向量，具体过程如下：

首先通过双向 LSTM 层得到每个时刻的输出并拼接，如公式（1）所述，接下来通过注意力机制计算每个时刻输入的权重，如公式（2）（3）（4）（5）所述。求得注意力层输出：

$$\boldsymbol{att} = contact\left(\boldsymbol{H}_1 * \alpha_1, ...\boldsymbol{H}_n * \alpha_n\right) \quad (9)$$

$$\boldsymbol{att} = \boldsymbol{W}_{att} * \boldsymbol{att} + \boldsymbol{b}_{att} \quad (10)$$

$\boldsymbol{att}$ 即为句子层输出的单个句子语义向量。

通过句子层的输出构建段落层的输入。获得每个句子的语义信息向量之后，可以认为这个集合足以表征整段文本的语义信息。由于本文的目的是提取文本中的切口数目，而每个句子中所描述的切口信息显然不平均，因此考虑先对句子层的输出进行信息过滤，注意此处仍使用注意力机制对每个句子的语义向量计算信息权重，用以实现一个信息门的过滤机制，此处亦与 HAN[25]有区别，这也是由于在医疗文本中，切口描述句之间有一定的独立性所决定的。此外，如前所述，一段文本中的某个切口描述句很有可能只是在重复描述前文的某个切口，即虽然整个句子的语义相对完整，句子间也相对独立，但是前后文之间的语义联系是客观存在的。考虑到这一点，本文在此采用 LSTM 机制，此处不同于 HAN[25]，并不采用双向 LSTM 机制是因为全文中句子之间的联系并不是太明显。因此构建段落层输入，每个时刻的输入为文本中相应的每个句子向量乘以其权重，获得段落层每个时刻的输出后再通过注意力机制计算权重，拼接降维，最后算出整段文本的语义向量，具体计算过程如下：

获得句子层的输出如下，设 $\boldsymbol{att}_i$ 表示第 i 个句子的输出向量：

$$\boldsymbol{Sen} = \left[\boldsymbol{att}_1, \boldsymbol{att}_2, ...\boldsymbol{att}_i, \boldsymbol{att}_{i+1}, ...\boldsymbol{att}_n\right]$$

通过注意力机制计算其权重，如公式 （2）（3）（4）（5）所述，求得其权重，利用加权进行信息过滤，构造段落层的输入如下。

$$\boldsymbol{Tex} = \left[\boldsymbol{att}_1 * \alpha_1, ...\boldsymbol{att}_i * \alpha_i, ...\boldsymbol{att}_n * \alpha_n\right]$$

将矩阵 $\boldsymbol{Tex}$ 输入 LSTM 层，每个时刻的输入为矩阵的每个向量，获得 LSTM 层每个时刻的输出，按照公式（2）（3）（4）（5）计算注意力权重，并按照公式（9）（10）所述的具体注意力机制，算得段落层最终输出作为文本的语义向量，并通过全连接层降为 4 维，用以表征四个分类的可能性，如公式（6）

（7）（8）所述，最终通过 softmax 处理得出最终切口数目预测值。

考虑到本模型多次使用全连接层降维，参数较多，因此降低了 LSTM 节点个数。基于 word2vec 训练模型得到文本中各词的词向量 100 维，双向 LSTM 隐含层维度以及 LSTM 层隐含层维度均设置为 32 维，为提高速率采用批量训练，batch_size 大小设置为 32，dropout 概率设置为 0.5，最终通过 1 个全连接层降维，最后的输出维度为 4，对输出向量进行 softmax 激活函数操作，取出概率最大的序号作为每段文本分类结果。

4 实验与分析

4.1 实验设置

本文使用合作的医院方提供的手术病例报告进行实验，病例报告如前文所述描述了对病人进行的手术处理以及检查和手术结果，其中包括手术过程中产生的切口种类、切口位置、操作流程等与切口相关的信息以及手术过程中的其他操作和术后处理。

通过对合作医院提供的医疗数据进行整理，最终得到手术报告 3000 例，通过人工对整段手术病例文本进行标注切口数目以及对“切口”抽取句进行切口数目标注后分别以此作为输入数据用于训练模型，采用交叉验证的方法进行实验，每次实验训练集、验证集、测试集划分比例为 3:1:1。

本文采用 Jieba[2]分词系统对文本进行分词。采用 word2vec 模型中的 CBOW 模型和 Negative Sampling 解法，将分词后的文本作为数据集训练词向量模型。Word2vec 模型在词向量训练中的应用十分广泛，它能够保证单词转换为词向量之后在向量空间中还能够维持词语之间的语义相似度。并且解决了词向量维数维度爆炸的问题。设置 word2vec 训练神经网络隐藏层维数大小为 100，是因为文本数量并不是特别大，且在实验过程中 100 维的效果较好。

4.2 对比模型

在本节中本文首先对第三节 3.1 小节句子切口数目提取模型进行了切口预测实验，同时为了体现对比效果，也记录了一些中间模型的实验效果作为对比模型如下所示；其次对 3.2 小节层次化切口抽取模型进行了实验，由于层次化模型基于分句模型，本文将分句模型中效果最好的双向 LSTM 与 attention 机制迁移过来，此时不再进行多个对比模型实验。

全文+LSTM：不进行分句，将整段病例文本分词并转换词向量后作为输入，通过 LSTM 机制构建整段病例文本的分类模型。

全文+BiLSTM：在上述模型基础上，将 LTSM 机制改为双向 LSTM 机制，通过双向 LSTM 机制构建整段病例文本的分类模型。

全文+LSTM+Attention：在全文+LSTM 模型基础上增加 attention 机制，通过 LSTM 神经网络与 attention 机制结合构建整段病例文本的分类模型。

全文+BiLSTM+Attention：在全文+LSTM+attention 模型的基础上，将 LSTM 机制改为双向 LSTM 机制通过双向 LSTM 神经网络与 Attention 机制结合构建整段病例文本的分类模型。

全文+SVM：不进行分句，将整段病例文本作为语料进行分词以及词向量训练。观察病例文本发现手术病例中对“引流管”的操作与切口操作类似，因此抽取其中与“切口”相关性高且与“引流管”相关性低的最佳 100 个词作为特征词，对文本进行特征词词频统计作为文本特征向量输入，训练 SVM 模型进行分类预测。

全文+TextCNN：不进行分句，将整段病例文本分词并转换词向量后作为输入，基于 CNN 神经网络构建模型，采用一个卷积池化层以及两个全连接层降维得出最终结果。

句子+LSTM：进行分句，提取切口描述句分词并转换词向量后作为输入，模型结构和全文+LSTM 中相同，最后结果进行累加。

句子+BiLSTM：进行分句，提取切口描述句分词并转换词向量后作为输入，模型结构和全文+BiLSTM 模型中的模型相同，最后结果进行累加。

句子+LSTM+Attention：进行分句，提取切口描述句分词并转换词向量后作为输入，模型结构和全文+LSTM+attention 中的模型相同，最后结果进行累加。

句子+BiLSTM+Attention：进行分句，提取切口描述句分词并转换词向量后作为输入，模型结构和全文+BiLSTM+Attention 中的模型相同，最后结果进行累加。

句子+SVM：进行分句，将所有切口描述句作为语料进行分词和词向量训练。模型结构和全文

[2] Jieba. https://github.com/fxsjy/jieba

表 2 实验结果

（a）不同分词的分词精度以及实验效果

方法	acc.	macro F1	macro pre.	macro recall	分词精度
jieba 分词	98.1%	98.0%	98.1%	97.2%	78.5%
合并词组	98.3%	97.6%	98.2%	96.8%	86.9%

（b）最终切口数目预测模型实验结果

方法	acc.	macro F1	macro pre.	macro recall
全文+				
LSTM	83.6%	43.3%	41.7%	45.2%
LSTM+Attention	93.3%	69.1%	69.8%	68.4%
BiLSTM+Attention	84.5%	51.3%	59.7%	55.4%
BiLSTM(TextRNN)	85.1%	52.9%	62.1%	51.2%
SVM	83%	43.8%	45.5%	43.0%
TextCNN	77.1%	40.1%	41.6%	40.1%
句子+				
LSTM	91.2%	91.1%	93.3%	91.2%
LSTM+Attention	96.3%	95.6%	96.1%	95.0%
BiLSTM+Attention	**98.1%**	**98.0%**	**98.1%**	**97.2%**
BiLSTM(TextRNN)	95.8%	95.1%	95.4%	94.7%
SVM	60.8%	17.4%	59.2%	23.4%
TextCNN	61.2%	44.5%	48.7%	45.2%
HAN	90.4%	84.5%	80.2%	89.3%
层次化切口提取	**98.6%**	**97.8%**	**97.7%**	**98.1%**

+SVM 中的模型相同，最后结果进行累加。

句子+TextCNN：进行分句，将切口描述句分词并转换词向量后作为输入，模型结构和全文+TextCNN 中的相同，最后结果进行累加。

本文测试了各种模型的预测准确率 accuracy、F1 值宏平均、precision 值宏平均以及 recall 值宏平均作为评价指标，同时给出了预测结果示例和置信度。

4.3 实验结果分析

由于本文以词向量为基础构建模型，因此在此首先讨论分词精度对于模型结构的影响。首先直接使用 jieba 中文分词工具对医疗文本进行分词，其次在此基础上再采用 word2vec 工具包中的 word2phrase 工具输出一次合并词组后的分词结果（实验证明一次合并效果最好），比较两种分词效果精度以及影响模型最终结果如下（表 2 中(a)）。对于分词效果精度，是通过随机抽取 100 例文本人工计算其分词精度来表征，对于模型最终结果，此处用的是层次切口抽取模型的结果（性能较好）。接下来本文将具体介绍模型结果。

4.3.1 模型结果对比

首先从表 a 中可以看到,合并词组之后，分词精度有了一定的提升，除此之外，在准确率以及其他的评价指标上，并没有明显提升，因此可以认为分词误差对实验结果的影响并不大。

接下来本文首先对句子切口数目提取模型进行实验，为了研究分句处理方法的有效性，提供不分句进行实验的全文输入对比模型；此外为了研究各种分类机制的效果，对 4.2 节所描述的各种对比模型进行实验，每种分类方法进行了交叉测试。随后由于本文层次化模型是在分句模型基础上提出的，并将分句模型中效果最好的神经网络机制迁移过来，因此对于层次化模型不再进行各种机制的对比，而是直接比较其模型效果与分句模型效果，研究其在保证模型整体性的基础上是否能同时保证分类效果，同时对比层次化切口抽取模型与 HRN，验证本文的机制更为有效。实验结果如表 2（b）所示。

（1）本文模型与其他分类模型的对比

通过表 2（b）可以看到，**本文提出的模型（句子+BiLSTM+Attention 以及层次化切口提取模型）取得了非常良好的切口数量抽取效果，整体抽取精度高达 98%**。此外，基于双向 LSTM 和注意力机制的模型，无论是在整段文本分类还是分句分类时的准确率和 F1 宏平均值都明显高于传统的 SVM 模型、TextRNN 以及 TextCNN 模型，且在病历文本上的最终切口数量判定结果上也都更好。在本文所使用的实验数据中，切口数目为 0 和为 3 的文本数量较多，文本分布不平均，通过 F1、precision 以及 recall 值可见 SVM 以及 CNN 的模型容易受到数据倾斜的影响。实验结果表明，对于短文本分类，深度学习的方式在准确率方面和相对数据模型的稳定性方面比传统的机器学习方式更有优势。而且相比于 SVM 模型，LSTM 模型不需要手动提取特征，无需加入核函数，并且可以以较高的速率处理大量的数据，效率较高且效果更好。所以就自然语言处理方面的分类来看，LSTM 神经网络更加具有说服力。而针对本文的文本分类问题，LSTM 循环神经网络比卷积神经网络

CNN 效果更好。

（2）分句分类与整个病例文本输入分类

从准确率的角度来看，比较表 2（b）中整段文本输入方式和句子输入的方式，整体上对比四种模型，**由句子作为输入比整段文本作为输入的效果更好**，最高可将准确率提升 14%（在使用双向 LSTM+Attention 时）；最低准确率提升了也有 3%（在使用 LSTM+Attention 时）。究其原因，本文认为对于表 2（b）中准确率相对不高的整段输入模型实验（（b）中全文+LSTM 模型、全文+BiLSTM+Attention 模型、全文+BiLSTM 模型）而言，由于是将整个段落文本作为输入，而整段手术报告中并不全是关于切口的描述，包含了很多其他与切口无关的手术处理（如前文所述），这属于冗余信息。而上述模型对于输入序列包含的所有信息都要进行处理，并没有冗余信息的除杂操作，因此实验结果容易受到冗余信息的干扰。而使用短句作为输入时，把文本中并不是对切口进行描述的内容去掉了，去掉了冗余的信息，从而使得输入最大限度地包含了有效信息，上述的三个模型在这种输入下进行训练与预测（句子+LSTM、句子+BiLSTM、句子+BiLSTM+Attention），结果比包含大量冗余信息的情况，显然要更稳定且准确率更高。

而对于其中加入了 Attention 机制模型，在整段文本输入的情况下（全文+LSTM+Attention），准确率已经可以达到 93%，从一个角度说明了本文注意力机制的有效性。而在句子输入的情况下（句子+LSTM+Attention），虽然准确率较全文输入有一定的提升，但是只有 3%。分析其原因，Attention 机制意味着模型对切口描述部分的关注度会更高，这相当于一个冗余信息过滤的机制，所以即使以包含大量冗余信息的文本作为输入，该模型可以自动提取有效信息并基于此进行训练与预测，因此在这种模型实验下，分句输入的优化空间不是太大，故在 Attention 模型中分句输入与整段文本输入相比的效果提升相对不如其他实验明显。

而从数据稳定性的角度来看，针对句子的分类模型相比于全文分类模型，其 F1 值、presicion 值以及 recall 值有了非常明显的提升。一方面，是因为将文本切分单独对句子进行考虑时，每个切口数目类别(0 个、1 个、2 个、3 个)的句子数目相对文本数目变得更为平均，数量也更充分，数据质量上有一定的提升；另一方面则是因为针对句子训练的模型能够更准确地把握句子语义从而判断其中描述的切口数目，模型效果更好，因此不易受到数据偏斜的影响，性能更好。

表 3 attention 效果

表 1 中例④、⑤、⑥权重词

例④词	权重	例⑤词	权重	例⑥词	权重
操作	0.1486	第	0.0826	吻合	0.0893
切口	0.1283	7	0.0813	右	0.0596
小	0.0818	肩胛	0.0799	间断	0.0290
做	0.0719	肋间	0.0738	支气管	0.0198
置入	0.0454	线	0.0644	积液	0.0183
于	0.0315	第	0.0639	胸腔	0.0174
胸腔镜	0.0308	5	0.0501	主	0.0170
肋间	0.0256	肋间	0.0486	水	0.0167
胸腔镜	0.0253	及	0.0433	3cm	0.0165
切口	0.0245	前线	0.0370	肋间	0.0164

综上，对于病历文本输入的预测，**分句处理比整段文本直接分类的准确率和稳定性都要更好，这说明了针对句子分析的有效性**。

（3）Attention 机制模型效果分析

表 2（b）显示，整段文本输入时，**使用 Attention 的模型（全文+LSTM+Attention）明显要优于对应的不使用 Attention 的模型(全文+LSTM、全文+BiLSTM)**，提升到了 93%，如前所述，其原因主要是因为 Attention 能够从文本中有效识别出对分类信息有用的关键信息。另外，如表 2（b）中句子输入的情况，也可以看到增加注意力机制之后（句子+LSTM+Attention、句子+BiLSTM+Attention）效果也有一定的提升。

但是，可以很明显地看到，整段文本输入时，用双向 LSTM 替换 LSTM 并加上 Attention 机制之后（全文+BiLSTM+Attention），准确率不升反降，从 93%掉到 84%，考虑其原因，本文认为，由于双向 LSTM 机制是将后文信息与前文信息综合，在整段文本输入的情况下，一个句子往往是描述一个完整的操作，句子与句子间的相关性不高，后文信息与前文信息的关联并不紧密，所以并不是很适合双向 LSTM 机制，这一点从全文+LSTM 模型与全文+BiLSTM 模型的比较也可以看出，单纯使用 LSTM 的模型效果与只使用双向 LSTM 模型的效果并无提升。而此时由于段落文本中冗余信息较多，使用双向 LSTM 机制将前后文信息综合，可能

增大冗余信息对注意力机制的干扰，所以准确率不升反降。

再从数据稳定性的角度分析，增加了Attention机制的模型，除了在双向 LSTM 结合的情况以外（前面已分析过原因），模型的 F1 值、precision 值以及 recall 值都有一定程度的提升，尤其是整段文本输入时增加 Attention 机制效果比较明显。

为了验证注意力机制的有效性，即 Attention 机制是否能够正确地对切口描述部分提高关注度，本文选取了一些医疗文本输出其在 Attention 机制下关注度权重最高的十个词，权重词提取源文本如表 1 中④、⑤、⑥的几个切口描述句所示，权重词提取结果如表 3 所示。

对于表 1 中例④，可以看到模型准确地找到了文本中描述切口的句子，并且抽取了相关的词语，比如“切口”一词，以及与切口位置相关的“肋间”一词，以及与操作相关的“置入”一词；对于表 1 中例⑤，可以看到由于文本中有的“切口”比如“延长右侧第 3 肋间腋前线小切口”并不表示手术中真的新开了切口，而是对之前已有切口的操作，所以模型没有把“切口”当作高权重词，而是把对切口的位置进行描述的词语抽取出来，并且准确地找到了相隔较远的两个术中切口描述语句，比如“第七肋间”、“第五肋间”以及“肩胛线”；对于表 1 中例⑥，可以看到在这段文本中，实际上是没有切口操作的，所以对于描述性的“切口”一词，模型并没有将其作为高权重词，模型输出了一些与切口操作无关的词，判定结果从而是零切口，虽然模型最后几个词实际上找到了文本中和切口操作描述相近的地方，比如描述肿物直径的词句，但是权重非常低。验证了本文的注意力机制有效性。

综上，Attention机制在准确率和数据稳定性的提升上是有意义的，且通过表 3 可以看到关于切口的部位大部分能够准确被找到，可以后续考虑通过 Attention 机制提取切口部位。因此在整段文本作为输入的情况下，使用LSTM构建模型以及增加Attention（全文+LSTM+Attention）是最好的选择，而在分句输入的情况下双向 LSTM 与 Attention 机制的结合（句子+BiLSTM+Atttention）效果最好。

（4）双向 LSTM 模型机制效果分析

分析表 2（b）中分句输入的模型，对于分句形式输入的情况，在输入的这一步已经将冗余信息进行了一定程度的过滤，所以有效信息的过滤不再是瓶颈，因此其中所有分句模型准确率都在90%以上，增加Attention机制效果较整段文本输入的情况虽不太明显，但也有一定程度的提升。但是可以看到在表 2（b）分句输入的模型中，双向 LSTM 机制（句子+BiLSTM、句子+BiLSTM+Attention）表现良好，针对句子分类的准确率可以达到 95%以上，同时数据稳定性也很好，F1 值、precision 值以及 recall 值平均都在 95%以上，究其原因，本文认为，对于整个句子来说，前后信息联系紧密互相影响，并且句子当中有效信息集中，综合前后文信息不会造成更多无用信息干扰，比较适合双向LSTM机制，所以使用双向 LSTM 机制替换 LSTM 机制的模型效果会更好，而对于整个文本，双向 LSTM 机制并不适用，这为层次化模型的机制选择提供了实验基础。

综上，在句子输入的情况下，使用双向 LSTM 构建于 Attention 机制综合模型（句子+BiLSTM+Attention）是较好的选择，并且是 4.2 小节中所有对比模型中效果最好的模型。

（5）层次化切口提取模型

如前文所验证的分句机制的有效性，层次化的模型实际上也有分句处理的机制，在句子层采用的是分句实验中效果最好的双向 LSTM+Attention 机制，在段落层根据全文比较适合 LSTM 的结论采用 LSTM+Attention 机制。可以看到实验中层次化切口提取模型的确在各个指标都超出了不分句的全文输入模型，而同时可以看到，在与分句模型进行对比时（句子+BiLSTM+Attention），层次化模型能够达到同样的良好的分类效果，在准确值 accuracy、召回率 recall 上有提升。而正如前面所分析的，层次化模型是端到端的，能够保证模型学习的完整性，同时并不需要像分句模型一样再次进行句子级别的标注，也不需要设计简单提取规则来筛选切口描述句，因此将模型迁移到其他问题也能有良好的拓展性。同时对比 HRN 的效果可以看到，本文的切口模型在各个指标上都要更优越。综上，本文基于分句模型提出的层次化切口提取是有效的。

（6）最优切口数目提取模型

综上，本文认为，提取手术报告中的切口数目问题可以转化为分类问题，本文在第 3 节提出的模型也就是在分句输入基础上，**构建双向 LSTM 与注意力机制结合最后进行汇总的模型（句子+BiLSTM+Attention）在这个问题上表现良好。而基于这个思路进一步提出的层次化切口抽取模型在分类效果良好的同时，具有更强的可移植性。**这

个模型综合考虑了文本的句子与段落之间的层次性，提出了分层处理的办法。并且由于句子内部语义联系紧密，在句子层使用双向LSTM比传统的LSTM 效果更优，而句子之间联系相对不紧密但存在信息传递，因此首先对句子语义向量进行信息过滤再通过 LSTM 层处理。同时使用 Attention 机制，使得模型对于预测切口数目的关键语句部分有准确的把握，对病历文本的切口数目的预测准确率可达到 98.4%。

5 结论

本文针对手术病例中切口数量的抽取问题进行了深入研究。巧妙地将切口数量抽取问题转换为病例内部句子的文本分类问题，基于双向 LSTM 机制构建了用于句子中切口数目自动提取的模型，同时进一步构建了层次化切口抽取模型。实验表明这两种方法都能够有效识别出病例中的切口数量，抽取准确度达 98%，而后者同时具备可拓展性强的优点。

本文以切口数量抽取问题为起点，对医疗文本中的信息抽取问题进行了初步探索。下一步我们将探索更多的不同类型数据的抽取方式，如手术病例中具体的切除部位、切除范围和出血量等，并探索和总结医疗文本中信息抽取的通用方法。

参 考 文 献

[1] Li Chang-Feng, Ke Si-Si, Liu Xin-Hui, Yan Ya-Qiong, Li Fang, Wang Liang . Seasonal Trends of Inpatient Number and Hospital Expenses in Primary Health Care Facilities. Chinese Journal of Social Medicine, 2017, 06(34):608-611(in chinese)
(李长风, 柯思思, 刘新会, 严亚琼, 李芳. 基层医疗机构住院量和住院费的季节性与变化趋势研究, 中国社会医学杂志, 2017, 06(34):608-611)

[2] Stacey L. Slager, Charlene R. Weir, Heejun Kim, Javed Mostafa, Guilherme Del Fiol. Physicians' perception of alternative displays of clinical research evidence for clinical decision support – A study with case vignettes. Journal of Biomedical Informatics, 2017, 71(2017):53-59

[3] Xu Ran. The research of intelligence auxiliary disease guidance based on text mining technology [Master thesis]. Beijing University of Posts and Telecommunications, Beijing, 2015. (in chinese)
(徐冉. 基于文本挖掘的疾病辅助导诊技术研究[硕士学位论文], 北京邮电大学, 北京, 2015)

[4] Liu Li-Ming. Risk modeling of cardiovascular disease risk factor discovery and early warning based on data mining [Master thesis]. Shenzhen University, Shenzhen, 2017. (in chinese)
(刘利明，基于数据挖掘心血管疾病风险因子发现与早期预警的风险建模 [硕士学位论文]，深圳大学，深圳，2017)

[5] Ni Xiao-Hua, Information Extraction of Non-structured Electronic Medical Records. China Digital Medicine, 2016, 11(12):89-91.(in Chinese)
(倪晓华. 非结构化电子病历信息的抽取. 中国数字医学, 2016, 11(12):89-91)

[6] Ruan Tong, Gao Ju, Feng Dong-Lei, Qian Xi-Yuan, Wang Ting, Sun Cheng-Lin, Process and methods of clinical big data mining based on electronic medical records. Big data research, 2017, 3(5):201754-doi:10.11959/j.issn.2096-0271.2017054(in Chinese)
（阮彤, 高炬, 冯东雷, 钱夕元, 王婷, 孙程琳. 基于电子病历的临床医疗大数据挖掘流程与方法. 大数据, 2017, 3(5): 2017054-doi:10.11959/j.issn.2096-0271.2017054)

[7] Li Zhou, Joseph M Plasek, Lisa M Mahoney, Neelima Karipineni, Frank Chang, Xuemin Yan, Fenny Chang, Dana Dimaggio, Debora S. Goldman, Roberto A. Rocha. Using medical text extraction, reasoning and mapping system (MTERMS) to process medication information in outpatient clinical notes. Proceedings of the AMIA Annual Symposium . Washington, USA, 2011, 2011:1639-1648

[8] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, Christopher G Chute. Mayo clinical Text Analysis and Knoledge Extration System(cTAKES): architecture, component evaluation and applications. Journal of the American Medical Information Association, 2010, 17(5): 507-513.

[9] Maofu, Li Jiang, Huijun Hu. Automatic extraction and visualization of semantic relations between medical entities from medicine instructions. Multimedia Tools and Applications, 2017, 76(8):10555-10573.

[10] Andrew M. Redd PhD, Adi V Gundlapalli MD. PhD, Guy Divita PhD, Marjorie E. Carter MSPH, Le-Thuy Tran PhD, Matthew H. Samore MD. A ploit study of heuristic algorithm for novel template identification from VA electronic medical record text, Journal of Biomedical Informatics, 2017, 71(2017):68-76

[11] Kevin Buchan, Michele Filannino, Özlem Uzuner. Automatic prediction of coronary artery disease from clinical narratives. Journal of Biomedical Informatics, 2017, 72(2017):23-32

[12] Roy J. Byrd, Steven R. Steinhubl, Jimeng Sun, Shahram Ebadollahi, Walter F. Stewart. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. International Journal of Medical Informatics, 2014, 83(12):983-992.

[13] Ayush Singhal, Michael Simmons, Zhiyong Lu. Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. Journal of the American Medical Informatics Association, 2016, 23(4):766-772.

[14] Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, Ulf Leser. How to improve information extraction from German medical records. It- Information Technology, 2017, 59(4):171-179.

[15] Sunghwan Sohn, Cheryl Clark, Scott R Halgrim, Sean P Murphy, Christopher G Chute, Hongfang Liu. MedXN: an open source medication extraction and normalization tool for clinical text. Journal of the American Medical Informatics Association, 2014, 21(5):858-865.

[16] Dai Bin-Rong, Wang Xiao-Li, Li Chao, Chen Jie, Shi Tian-Xing, A mining and classification method for medical data based on PCA-SVM. Computers Applications and Software, 2016, 33(8) :67-70 (in Chinese)
（戴炳荣，王晓丽，李超，陈浩，施天行．一种基于PCA-SVM的医疗卫生数据挖掘分类方法．计算机应用与软件，201, 33(8) :67-70)

[17] Xu Teng. Research of Clinical Data Mining and Analysis Based on Thyroid Disease. [Master thesis]. Donghua University, Shanghai, 2016. (in chinese)
(许腾．基于甲状腺疾病的临床数据挖掘与分析研究[硕士学位论文]，东华大学，上海, 2016)

[18] Nie Bin, Wang Zhuo, Du Jian-Qiang, Zhu Ming-Feng, Lin Jian-Ming, Ai Guo-Ping, Xiong Lin-Zhu,The Study on Classification of Secondary Complications of Diabetes Based on Rough Set and Random Forest. Journal of Jiangxi Normal University(Natural Science Edition), 2014, 03(38): 278-281 (in Chinese)
(聂斌，王卓，杜建强，朱明峰，林剑鸣，艾国平，熊玲珠．基于粗糙集和随机森林辅助糖尿病并发症分类研究．江西师范大学学报, 2014, 03(38):278-281)

[19] Gong Fan, Wang Meng-Jie, Ruan Tong, Wang Hao-Fen, Lu Hao, Automatic Recognition Methods of Symptoms in Texts of Electronic Medical Records. Journal of Medical Informatics, 2016, 37(7):7-14.(in Chinese)
(龚凡，王梦婕，阮彤，王昊奋，陆灏．电子病例文本症状自动识别方法．医学信息学杂志, 2016, 37(7):-14)

[20] Yang Zhi-Hao, Hong Li, Lin Hong-Fei, Li Yan-Peng, Extraction of information on protein-protein interaction from biomedical literatures using an SVM. CAAI Transaction on Intelligent Systems, 2008, 3(4):361-369.(in Chinese)
(杨志豪，洪莉，林鸿飞，李彦鹏．基于支持向量机的生物医学文献蛋白质关系抽取．智能系统学报, 2008, 3(4):361-369)

[21] Yoon Kim. Convolutional Neural Network for Sentence Classification. Proceeding of the 2014 Conference on Empirical Methodes in Natural Language Processing. Doha, Qatar, 2014, 1746-751.

[22] Pang Liang, Lan Yan-Yan, Xun Jun, Guo Jia-Feng, Wan Sheng-Xian, Cheng Xue-Qi. Review of depth-text matching. Chinese Journal of Computers, 2017, 40(4): 985-1003 (in Chinese)
(庞亮，兰艳艳，徐君，郭嘉丰，万圣贤，程学旗．深度文本匹配综述．计算机学报, 2017, 40(4): 985-1003)

[23] Bahdanau D, Cho K, Bengio Y, Neural machine translation by jointly to align and translate. Proceedings of the ICLR 2015, San Diego, USA, 2015: 1-15.

[24] Minh-Tang Luong, Hieu Phan Christop D.Maning, Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methodes in Natural Language Processing. Lisbon, Portugal, 2015, 1412-1421.

[25] ZichaoYang, DiyiYang, ChrisDyer, XiaodongHe, AlexSmola, EduardHovy. Hierarchical Attention Networks for Document Classisication. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016, 1480-1489.

[26] Zaiqing Nie, Jirong Wen, Weiying Ma. Statistical Entity Extraction From the Web. Proceedings of the IEEE. 2012, 100(9):2675-2687.

[27] Chunyu Yang, Yong Cao, Zaiqing Nie, Jie Zhou, Jirong Wen. Closing the Loop in Webpage Understanding. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(5):639 -650.

Lu Shuqi ,born in 1997,master candidate. Her research interests include Natural Language Processing and Data Mining.

Dou Zhicheng, born in 1980, Ph.D., professor. His research interests include Information Retrieval, Data Mining, and Big Data Analytics.

Wen Ji-Rong, born in 1972, Ph.D., professor. His research interests include Information Retrieval, Database, Data Mining, and Big Data Analytics

Background

The problem discussed in this paper is related to the fields of text extraction and data mining. Text extraction and text information structuring has become a research hotspot in the medical field. With the help of extraction of structured information from medical texts, the doctors can effectively save time and energy on reading and understanding the long text, and spend more time on diagnosis and treatment plan for the patient. But if only rely on manual extraction of structure information from the surgical cases, it will consume a lot of manpower and material resources.

An effective method for solving this problem is to design rules based on regular expression or using entity extraction algorithm based on CRF to realize automatic text extraction. A common problem of these algorithms is that the method of rule matching can only extract the key information in the text, and it is difficult to quantify the information, and for complex case description text, text matching rules are difficult to design.

In order to solve this problem, we take the extraction of the number of incisions in thoracic surgical cases as example, using the idea of text classification to extract the number of incisions. Through our classification model based on deep learning constructed in this paper, we find that the accuracy of extracting the number of medical text incisions can reach 98 %, and also has better performance in terms of data stability than those methods based on rule designing or those traditional SVM-based classification models.

The authors of the paper have done lots of research on text extraction and natural language processing, like Web entity extraction (please refer to[26]). They proposed a novel framework called WebNLP which enables bidirectional integration of page structure understanding and text understanding in an iterative manner in IEEE TKDE (please refer to [27]).

This work was supported by the National Natural Science Foundation of China (Grant No. 61872370) and the National Key R\&D Program of China (No. 2014CB340403).